Regulatory Toxicology and Pharmacology 72 (2015) 458–472

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Empirical characterisation of ranges of mainstream smoke toxicant yields from contemporary cigarette products using quantile regression methodology

Oscar M. Camacho *, Alison Eldridge, Christopher J. Proctor, Kevin McAdam

Group Research and Development, British American Tobacco (Investments) Ltd, Southampton, UK

ARTICLE INFO

Article history:
Received 19 December 2014
Available online 27 May 2015

Keywords:
Cigarette smoke toxicants
Toxicant yields
N′-nitrosoanatabine
Acetone
Pyridine
Phenol
Tobacco regulation
Quantile regression
Percentiles

ABSTRACT

Approximately 100 toxicants have been identified in cigarette smoke, to which exposure has been linked to a range of serious diseases in smokers. Smoking machines have been used to quantify toxicant emissions from cigarettes for regulatory reporting. The World Health Organization Study Group on Tobacco Product Regulation has proposed a regulatory scenario to identify median values for toxicants found in commercially available products, which could be used to set mandated limits on smoke emissions. We present an alternative approach, which used quantile regression to estimate reference percentiles to help contextualise the toxicant yields of commercially available products with respect to a reference analyte, such as tar or nicotine. To illustrate this approach we examined four toxicants (acetone, N′-nitrosoanatabine, phenol and pyridine) with respect to tar, and explored International Organization for Standardization (ISO) and Health Canada Intense (HCI) regimes. We compared this approach with other methods for assessing toxicants in cigarette smoke, such as ratios to nicotine or tar, and linear regression. We concluded that the quantile regression approach effectively represented data distributions across toxicants for both ISO and HCI regimes. This method provides robust, transparent and intuitive percentile estimates in relation to any desired reference value within the data space.

1. Introduction

Cigarette smoke is a dynamic and complex aerosol containing over 6000 identified components and possibly many thousands of further unidentified constituents (Perfetti and Rodgman, 2011). Approximately 100 harmful or potentially harmful compounds have been identified in cigarette smoke (US Food and Drug Administration, 2012), and exposure to these smoke constituents is believed to be responsible for a wide range of serious diseases amongst smokers (Fowles and Dybing, 2003; Rodgman and Perfetti, 2009). In the present work we focus on toxicants, chemical species in tobacco or cigarette smoke, exposure to which may result in harm to the tobacco user.

Observed health responses to toxicants are dependent on the intensity and duration of exposure, though dose–response relationships are only known through epidemiology and total exposure to cigarette smoke, and are generally not known for individual toxicants. The most widely accepted measures of exposure to cigarette smoke toxicants are biomarkers. However, relatively few validated biomarkers of exposure exist for individual cigarette smoke toxicants. Furthermore, biomarker measurement is invasive, slow and expensive; hence limited data are available on their levels in smokers (Hatsukami et al., 2003; Hecht et al., 2010) and few inter-laboratory comparisons have been made of these data (Minet et al., 2011). Consequently, their utility in understanding smokers' exposure to toxicants is somewhat restricted given current scientific capabilities in this area. In recent years, mouth-level exposure approaches have been developed that examine used cigarette butts to estimate individual human exposure to nicotine, nicotine-free dry particulate matter (NFDPM, tar) and a small number of individual toxicants. This approach shows

Abbreviations: BAT, British American Tobacco; CF, Cambridge filter; CORESTA, Cooperation Centre for Scientific Research Relative to Tobacco; CV, coefficient of variation; dwb, dry weight basis; FCTC, Framework Convention on Tobacco Control; FTC, Federal Trade Commission; GC–MS (EI), gas chromatography with mass spectrometry using electron impact ionisation; HCI, Health Canada Intense; ISO, International Organization for Standardization; NAT, N′-nitrosoanatabine; NFDPM, nicotine-free dry particulate matter; ppm, parts per million; TobReg, WHO Study Group on Tobacco Product Regulation; TPM, total particulate matter; TSNA, tobacco-specific nitrosamines; WHO, World Health Organization.

* Corresponding author at: Regents Park Rd., Southampton SO15 8TL, UK. Tel.: +44 (0) 2380 588 258.

E-mail addresses: Oscar_M_Camacho@bat.com (O.M. Camacho), Alison_Eldridge@bat.com (A. Eldridge), Christopher_Proctor@bat.com (C.J. Proctor), Kevin_McAdam@bat.com (K. McAdam).

http://dx.doi.org/10.1016/j.yrtph.2015.05.023
0273-2300/

promise, but it is limited to the set of toxicants within the scope of the technique.

Historically, smoking machines have been used to quantify toxicant emissions from cigarettes (Baker, 2006). Several different regimes, or sets of smoking parameters, have been adopted for regulatory measurement and reporting of emissions. The general consensus is that smoking machine yields cannot predict actual exposure to cigarette smoke constituents in humans, because wide variability in smoking behaviour in any population will have a significant effect on toxicant exposure (US Department of Health and Human Services, 2001). However, the machine smoking approach enables standardised measurement (International Organization for Standardization, 2000) and provides an established platform for comparing emissions from different products. Some scientific and regulatory groups have proposed using two regimes as a means of estimating the lower and upper boundaries of possible emissions from cigarettes: the International Organization for Standardization (ISO) regime, which consists of a 35 mL puff of 2 s duration taken every 60 s (ISO 4387:2000) and a more intense regime developed by Health Canada Intense (HCI), which consists of a 55 mL puff of 2 s duration taken every 30 s and additionally all cigarette filter tip ventilation holes blocked using a strip of Mylar adhesive tape (Health Canada, 1999). Thus, despite deficiencies in relating machine measured yields to smokers' exposure, machine-based analysis of cigarette yields is likely to remain the prevalent method for quantifying and comparing toxicant emissions from cigarettes for some time to come (Hecht, 2012).

Smoking machines are used as the basis of regulatory reporting with regards to cigarette toxicant emissions in a number of geographic jurisdictions. Regulatory authorities in Brazil, Canada, Nepal, Taiwan, USA and Venezuela, have historically, or currently require measurement and reporting of toxicant emissions from cigarettes on sale in their jurisdictions. The World Health Organization (WHO), under its Framework Convention on Tobacco Control (FCTC) (WHO, 2005), is facilitating standardised approaches to tobacco regulation on a global scale. One of the initiatives under the FCTC is a working group, the WHO Study Group on Tobacco Product Regulation (TobReg), which recommends possible approaches to product regulation (Burns et al., 2008), has suggested an approach for measuring toxicants.

TobReg has proposed a regulatory scenario where every distinct cigarette product on a market is measured for a selective set of toxicants, and the data used to identify market medians, which could be used to set mandated limits on smoke emissions (Burns et al., 2008). Under this scheme, if products on sale in a market fail to meet these limits they would be prohibited. Limits are proposed for emissions of nine smoke toxicants (4-[methylnitrosamino]-1-[3-pyridyl]-1-butanone (NNK), N′-nitrosonornicotine (NNN), benzo[a]pyrene, formaldehyde, acetaldehyde, acrolein, 1,3-butadiene, benzene and carbon monoxide), expressed as ratios to nicotine measured under HCI smoking conditions. TobReg also suggest progressive reductions in the amounts of these toxicants in smoke over time, as technology becomes available to reduce them. In the USA, the Food and Drug Administration has required tobacco manufacturers to measure and disclose a larger number of individual toxicants and may in the future establish product standards, including ceilings on smoke emissions.

Despite the ongoing interest in mainstream smoke emissions and the number of data points reported to regulators on an annual basis, surprisingly few data have been published on toxicant yields from contemporary commercial cigarettes. Of the many thousands of cigarette brands on sale globally, ISO and HCI mainstream smoke yields have been reported for only around 150 (Australian Government Department of Health and Ageing, 2002; Counts et al., 2005; Gregg et al., 2004) (Tobacco Control Programme, Health Canada. Constituents and emissions reported for cigarettes sold in Canada—2004. Unpublished data received upon request from TRR_RRRT@hc-sc.gc.ca) and from a small geographical area (UK (Gregg et al., 2004), Australia (Australian Government Department of Health and Ageing, 2002) and Canada (Tobacco Control Programme, Health Canada. Constituents and emissions reported for cigarettes sold in Canada—2004. Unpublished data received upon request from TRR_RRRT@hc-sc.gc.ca)). Other extensive smoke yield data have been presented, but in ways that do not allow subsequent independent analysis. For example, Hyodo et al. (2007) published data in 2007 on Japanese cigarettes, where they presented ranges for toxicant yields and functional relationships with yield of tar, but did not provide individual yield data. Thus, although in the future it can be anticipated that a greater volume of smoke yield data will become available, the current dataset of machine yields for mainstream cigarette smoke is small, and there is no available contemporary picture of the range and diversity of toxicant levels generated by commercial cigarettes worldwide.

In an attempt to gain some insight into the range of toxicant yields of current commercially available cigarette products, we measured the toxicant emissions from a wide range of products over a number of years. The database currently consists of ISO smoke yields for 959 products, 364 for HCI smoke yields sourced from 80 geographical areas, and 916 blend chemistries (Supplementary Fig. 1).

The British American Tobacco (BAT) dataset includes cigarette products from a number of international and national manufacturers, and includes a range of cigarette formats (circumference, length, and filter type) and blend styles (flue-cured Virginia, US-blended, and blends disposed between these two styles). The database was assembled over the time period between 2007 and 2011. This dataset is of sufficient size to enable comparison of smoke emissions from different products and to characterise differences in smoke chemistry between many countries.

As a foundation for these analyses, a robust and standardised methodology for critical assessment of this type of data is required. We define robustness in this situation as the ability to estimate meaningful reference values from the data, but with these estimates showing little sensitivity to future incorporation of additional data in the database, or use of values at the extremes of the measured product ranges and/or anticipated levels of product variability over time.

As additional smoke yield data becomes available, an important question that arises is how best to analyse, understand and contextualise the range of toxicant levels and emissions from cigarettes. The predominant approach adopted to date has been on an individual per-product basis. Under this structure, toxicant precursor levels in cigarette blends are generally reported per gram of tobacco, either on a "dry-weight" basis (i.e. after correction for the moisture content back to a dry tobacco weight) or an "as-received" (wet-weight basis) value. Toxicant emissions from cigarettes are usually reported on a per-cigarette basis, although ratios of toxicants to nicotine under a specific smoking regime have also been proposed (Burns et al., 2008). Intrinsically, existing approaches for analysis of toxicants vary substantially and are dependent on the matrix in which they are measured and the way toxicants are reported. The methodology for data assessment should therefore respect the way the data is generated and reported. In addition, a framework with which to compare toxicant yields with global and historic values is likely to be of great value in contextualising and understanding smoke yields in the future.

In this article we explore a number of different approaches to analysing tobacco blend and smoke yield data. We examine univariate ratios, simple regression and quantile regression methodologies (Kroenger, 2005), to assess the toxicant precursor content and smoke toxicant yields of commercial cigarette products. The quantile approach uses prediction to estimate percentiles for a

specific feature with respect to other variables. It has been widely used in medicine to develop growth reference charts such as those commonly used to monitor babies height or weight versus age, (Kuczmarski et al., 2002; Wei et al., 2006), as well as in neurology (Benatar et al., 2009; Peng et al., 2009) and in clinical chemistry (McGreevy et al., 2009). Applications outside medicine are increasingly common, with uses seen in finance (Bouyé and Salmon, 2009), environmental sciences (Cade, 2011) and various other fields. In relation to research of tobacco smoke toxicants, quantile regression provides a framework for comparing the toxicant yields of any product with predicted percentile values (references) that have been calculated from other commercial products and with respect to tar yield. The complete dataset is used to calculate the predicted quantiles, and hence, it provides estimates that are more robust than those derived simply from analyses by interval.

2. Materials and methods

2.1. Data sources

We assembled a database of smoke toxicant yields from data published in the scientific literature (Counts et al., 2005; Gregg et al., 2004), regulatory reporting data (Australian Government Department of Health and Ageing, 2002; Tobacco Control Program Health Canada, 2004) and data from BAT's in-house analytical laboratories. Only three laboratories have generated these data: Arista Europe for the UK benchmark data (Gregg et al., 2004), Labstat International for the Australian, and Canadian and international data (Australian Government Department of Health and Ageing, 2002; Tobacco Control Program Health Canada, 2004) and BAT laboratories for all remaining data. Data for non-commercial products and duplicate values for commercial brands were removed from the database. Data were separated into ISO yields and HCI yields. Data derived from the USA Cambridge filter pad method (formerly known as the Federal Trade Commission [FTC] method) were not included with ISO values from other countries because small operational differences between the two smoking procedures could potentially lead to systematic differences in toxicant yields (Baker, 2002). The final database comprised data with measurements for 916 blends, 959 products measured under the ISO regime and 364 under HCI. There continues to be publications of toxicant datasets (Bodnar et al., 2012) which will allow recording of new entries in the database.

Data on smoke constituents can differ substantially between laboratories, especially for low-level smoke constituents, because of differences in approaches to the measurements employed by these laboratories (Purkis and Intorp, 2014). Additionally, year-to-year consistency within laboratories can vary notably even when there is consistency of methodology. Hence all measured smoke yield data is associated with a degree of operational inter-laboratory variability. However, taking data from only one laboratory would make the findings relevant to the output of that laboratory alone. Because we aimed to identify the extent of diversity in cigarette smoke toxicant yields from commercial contemporary cigarettes, it was essential to encompass the effects of these sources of variability. Therefore data from all available laboratories was included in this analysis.

We chose four cigarette smoke constituents to examine different approaches for summarising cigarette smoke toxicant data: N′-nitrosoanatabine (NAT) (a major tobacco-specific nitrosamine, sensitive to blend character and found in the particulate phase), acetone (a volatile carbonyl compound, sensitive to the inclusion of charcoal in cigarette filter), pyridine (a semi-volatile compound that demonstrates blend character sensitivity), and phenol (highly sensitive to the presence of a cellulose acetate filter and to blend chemistry). These compounds represent four of the main families of toxicants present in cigarette smoke. ISO data for all four compounds were examined, but for the HCI regime data we used NAT only, as this was sufficient to examine whether the same techniques explained for the ISO regime also applied to the HCI regime. We also reported the ratios of NAT to nicotine yields derived from the HCI machine smoking regime. NAT is a tobacco blend precursor to the smoke toxicant, and we examined its variety of yields in the 916 products. Descriptive statistics, including the number of products used in the statistical analysis for each endpoint are shown in Table 1.

2.1.1. BAT in-house analytical laboratory methods

BAT in-house analytical laboratories used multi-analyte methods to analyse cigarette mainstream smoke toxicant yields, whereby members of a group of toxicants (e.g. tobacco-specific nitrosamines, carbonyl or phenolic compounds) were analysed simultaneously from the same cigarette. Most of the methods followed, or were based on internationally accredited protocols (i.e. ISO, Cooperation Centre for Scientific Research Relative to Tobacco [CORESTA] or Health Canada official methods). All methods have been internally validated for repeatability and reproducibility (AOAC International, 2002; International Organization for Standardization, 1994).

Details of the analytical methods for the analytes reported here are contained in the Supplemental information. Details of all of the analytical methods used by the BAT laboratory for mainstream smoke toxicant emissions and cigarette filler blend analyses for the products contained in the database are available upon request.

2.2. Univariate and simple linear regression model approaches

Historically, categorisation of cigarette smoke yields has been based on univariate approaches and average values. We created

Table 1
Descriptive statistics for ISO, HCI, ratios and blend endpoints.

Analyte	Smoking regime	*N*	Mean (SD)	Median (min, max)
Tar (mg/cig)	ISO	959	8.7 (3.6)	8.9 (0.31, 20.8)
	HCI	363	27.1 (5.0)	26.7 (15.6, 40.7)
Nicotine (mg/cig)	ISO	959	0.73 (0.28)	0.72 (0.07, 1.8)
	HCI	363	1.96 (0.43)	1.93 (0.97, 3.67)
NAT in cigarette blend (ppm dwb)	n/a	916	0.92 (0.49)	0.87 (0.14, 3.72)
NAT (ng/cig)	ISO	959	58 (30)	54 (4, 194)
	HCI	364	143 (73)	129 (22, 414)
NAT/Nicotine (ng/mg)	ISO	959	86 (44)	84 (8, 406)
	HCI	363	76 (40)	76 (11, 204)
Acetone (μg/cig)	ISO	959	167 (72)	167 (0.6, 396)
Phenol (μg/cig)	ISO	959	13.5 (8.7)	12.8 (BLQ, 90.4)
Pyridine (μg/cig)	ISO	959	6.8 (4.1)	6.2 (0.2, 22.6)

BLQ is below limit of quantification; dwb, dry weight basis; n/a not applicable; ppm, parts per million; SD, standard deviation.

reference tables based on univariate data analyses and regression models with respect to tar, to aid comparisons with quantile regression outputs.

2.3. Quantile regression

Classic regression can provide references based on the mean at nominal tar levels. However, applicability of estimated yield from classic regression has several limitations in this context. Comparisons of real data to references are limited to assessing one value, the expected average value. Products with very different smoke yields could be equally categorised as being above or below the mean. Additionally, estimates based on means are sensitive to extreme values. Therefore updating the reference estimates using new observations could lead to very different estimates. Some attempts have been made to use the median as a reference (WHO, 2005), which is likely to yield more robust estimates because the median is not sensitive to the measurement scale, only to the relative position of those measurements. However, using only the median as a reference is not sufficient to represent the broad scope of smoke yields from commercial products. Quantile regression can be seen as an extension of the median approach, in which quantile regression lines can be used to estimate reference percentiles at nominal tar levels. For example, the quantile 0.5 can be used to estimate the references at the median. In quantile regression (Koenker and Basset, 1978), the τ th quantile of Y is defined as the inverse function $Q(\tau) = \inf\{y:F(y) \geqslant \tau\}$ where $0 < \tau < 1$. This function represents the value of the response variable for which the probability distribution function of y, that is $F(y) = P(Y \leqslant y)$, is larger or equal to the chosen τ. Quantile regression produces a model of explanatory variables or covariates (in this case tar) on the conditional percentiles of a response variable (specific toxicant yields).

A typical approach to regression uses the least-squares method to build a model defined by the conditional mean of the variable response Y given that the explanatory variable X has a particular value x. Similarly, quantile regression assesses the quantile function $Q(\tau \mid X = x)$. This conditional relationship can be written as the linear conditional function $Q(\tau \mid X = x) = x' \beta(\tau)$, which is resolved by minimising the objective function for $\tau \varepsilon (0,1)$ (1).

$$\beta(\tau) = \arg\min \beta \in R^p \sum_{i=1}^{n} \rho_\tau(y_i - x_i' \beta) \quad (1)$$

We took an empirical approach to choose the most appropriate model. Linear combinations of tar up to the cubic function were assessed: tar^{-1}, tar, $tar^{1/2}$, $tar^{1/2}$*tar, tar^2 and tar^3. Quantile regression modelling of polynomial relationships has been explored (Kim and Yang, 2011; SAS Institute Inc., 2011) and tables of estimates from a full model (2), a cubic polynomial model (3) and a quadratic polynomial (4) were compared:

$$\text{Toxicant} = \beta_{11} tar^{-1} + \beta_{12}\text{tar} + \beta_{13}\text{tar}^{1/2} + \beta_{14}\text{tar}^{1/2} * \text{tar} + \beta_{15}\text{tar}^2 + \beta_{16}\text{tar}^3 \quad (2)$$

$$\text{Toxicant} = \beta_{21}\text{tar} + \beta_{22}\text{tar}^2 \quad (3)$$

$$\text{Toxicant} = \beta_{31}\text{tar} + \beta_{32}\text{tar}^2 + \beta_{33}\text{tar}^3 \quad (4)$$

The classic cubic polynomial without intercept was chosen to create the final reference tables. The reasons for choosing this model over others are explained in Section 3.2.2.

Minimisation of the objective functions did not prevent the quantile curves crossing. To enable non-crossing regression, we applied the algorithm developed by Muggeo et al. (2013), chosen for its conceptual simplicity and ease of implementation. Broadly, linear constraints are introduced by calculation of quantile curves with respect to a starting quantile (such as the median (Muggeo, 2013)). We used a quantile regression package developed by Muggeo, the quantregGrowth package in R (Muggeo, 2013), which is based on the package Quantreg (Koenker, 2013), to facilitate estimation of curves. The package quantregGrowth also ensures monotone estimates across the explanatory variable, in this case tar. That is, toxicant levels increase or remain at the same level as the tar yield increases. This assumption is intuitively valid because the toxicant levels in smoke are expected to increase or approach saturation as the level of tar increases (Counts et al., 2005; Gregg et al., 2004; Health Canada, 1999; Hyodo et al., 2007).

To provide a measure of accuracy for the model's predictions, we calculated bootstrap confidence intervals (Carpenter and Bithell, 2000; Schall, 2012). This method consists of extracting numerous random samples with replacement of sample size equal to the original sample size (up to 959 observations for ISO and up to 364 observations for HCI, depending on the number of missing values for each toxicant). We carried out 10,000 iterations for 19 tar reference levels (1–19 mg) for ISO yields and eight tar reference levels (17–38 mg in 3 mg intervals) for HCI smoke yields. Predicted reference values were estimated for each percentile of interest on the basis of the median of each set of 10,000 values, with the bootstrap confidence intervals set at the 2.5th and 97.5th percentiles of the amalgamated dataset created with all predictions. Tables based on the 2.5th, 50th and 97.5th percentiles of the reference estimates were preferred over means and standard deviations because distributions of the bootstrap estimates did not seem to follow normality. Crossing quantile regression (without constraints) was carried out with PROC QUANTREG (Koenker, 2013) in SAS version 9.3, while non-crossing regression (with constraints) was carried out with the quantregGrowth package in R, version 2.12.0.

2.4. Data transformation and assumptions

An advantage of quantile regression over other possible approaches to characterise tobacco product smoke yields is that quantile regression does not make assumptions about the distribution of the data (Kroenger, 2005; McGreevy et al., 2009). Therefore, data transformation is not required and does not improve the fit of the regression. Some values were reported below the limit of quantification, which prevented them from being used in table computation. However, to omit these values would lead to a low bound censoring for tar, which could affect estimates for low tar products. To mitigate this effect, we imputed half the limit of quantification whenever it was known (i.e. reported), otherwise it was left as a missing value.

3. Results

3.1. Univariate analysis

3.1.1. Toxicant precursors in tobacco blends

Tobacco blends are complex mixtures of different types of tobacco, such as Virginia, Burley and Oriental tobacco, created as an amalgamation of cut leaf, stems, expanded and reconstituted tobaccos, as well as other ingredients and process aids. Within tobacco are compounds that are precursors to the smoke toxicants formed when the tobacco is burnt. Generally, the precursors for smoke toxicants are poorly characterised (Piade et al., 2013) and there are often multiple tobacco precursors for individual smoke toxicants (Baker et al., 1999). Nevertheless, significant correlations have been reported between some tobacco compounds and smoke toxicants. For example, strong links have been established between tobacco constituents and smoke toxicants for the tobacco-specific nitrosamines (Rodgman and Perfetti, 2009).

Toxicant precursor levels can vary both within and between tobacco varieties due to agronomic, geographic, environmental and temporal factors. Variability over time arises from factors such as weather, processing and storage. Therefore, analysis of a broad sample of products over an extended time period is essential to represent the expected diversity of toxicant content from commercial products. The cigarette tobacco filler blends analysed in this work included different blends from a range of manufacturers, sampled over the time period 2007–2011.

Percentiles are an intuitive and robust way to summarise the range of tobacco blend toxicants when analysing this type of dataset. Percentiles are intuitive, as they provide a set of reference values that facilitate comparisons with new samples, and robust, because for a well characterised dataset the percentile estimates do not change significantly as new data are introduced if the sampling space is well characterised.

Calculation of percentiles for the tobacco-specific nitrosamine NAT blend yielded measurements of 0.1–3.7 ppm (ppm) dry weight basis (dwb), with a median value of 0.9 ppm dwb (Fig. 1). The data were positively skewed, with most values <2 ppm dwb. Fig. 1 also displays the cumulative density function used for calculating percentiles at nominal points.

The data are summarised in Table 2 in the form of percentile estimates. Values for the 5th, 10th, 25th, 40th, median, 60th, 75th, 90th, 95th and 99th percentiles are shown in Table 2, but any reference point can be calculated.

Therefore, the univariate approach offers a practical way of summarising tobacco blend contents. In principle, further analysis and segregation of this kind could be performed on the dataset if sufficient information was available, e.g. blend style type, manufacturer, etc.

3.1.2. Cigarette smoke toxicant emissions

The same univariate data assessment approach can be used to provide percentile estimates for overall smoke toxicant yields across all products. Data are presented for ISO and HCI yields in Supplementary Fig. 2 and Table App1. These data provide a reference framework against which mainstream smoke yields from other cigarette products can be compared.

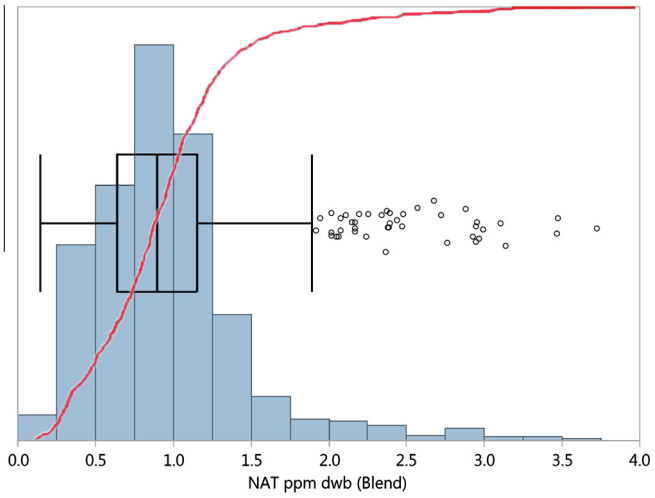

Fig. 1. NAT in blend distribution, boxplot and cumulative distribution. Where the histogram's *y*-axis represents the number of products which, divided by the total number of products, can be seen as an empirical distribution of this sample. The boxplot is formed by the median (central line); the box is the interquartile range, whiskers are situated at 1.5 times from quartiles; observations outside the whiskers (right side points) are often interpreted as extreme values. The red curve represents the cumulative density function (proportion of products with respect of the *x*-axis). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
NAT (ppm dwb) in blend percentiles.

Percentile	NAT ppm dwb (blend)
5th	0.324
10th	0.384
25th	0.641
40th	0.816
50th	0.896
60th	0.988
75th	1.150
90th	1.425
95th	1.779
99th	2.935

dwb, dry weight basis; ppm, parts per million.

However, inspection of the data revealed the distribution of smoke constituents to be highly dependent on levels of blend precursors (where known, e.g. blend NAT levels) and overall smoke yields (as represented by tar yield) from the products. Both of these parameters have been previously identified as contributing strongly to measured levels of toxicants. For example, Gregg et al. (2004), Counts et al. (2005) and Hyodo et al. (2007) all identified strong and positive correlations between the amount of ISO or HCI tar and the yields of individual smoke constituents. Correlations with species with identified blend precursors were strengthened by incorporation of parameters reflecting blend character or content.

The influence of overall smoke yield (as represented by tar yield) on the smoke yields measured in this study was examined (Fig. 2). Strong correlations with tar were observed for acetone, pyridine and phenol, with some outliers in the case of phenol for both ISO and HCI regimes. Weaker but positive and significant correlations between tar levels and NAT yields were also identified.

Comparison of the data for acetone, NAT and pyridine across the two smoking regimes showed continuity against tar, irrespective of regime. This was not the case for phenol, which indicated a down shift in the data from ISO to HCI measures. Toxicant data did not fully overlap between the two smoking regimes, as there was an area of scarce data between 16 mg and 20 mg tar. Given that there were cases of discontinuity between regimes and scarcity of data in the joining region, we assessed ISO and HCI yields separately.

We also examined the impact of blend NAT levels on smoke NAT yields. The data were plotted as tar normalised NAT yields against blend NAT levels for both ISO and HCI regimes (Fig. 3). The Spearman's correlation coefficient for NAT in blend and mainstream smoke from cigarettes smoked under the ISO regime was $R = 0.5615$, whereas that between NAT in blend and yields measured using the HCI regime was $R = 0.7810$.

Clearly, the univariate approach for summarising and categorising mainstream smoke yields is a relatively crude simplification of the observed variation in cigarette smoke yields, and the reasons underlying the range of values. As noted above, blend toxicant precursors are poorly characterised and rarely measured. In contrast, parameters such as tar, nicotine and carbon monoxide are routinely measured for cigarettes, as they are often used as quality control and design parameters for cigarette products, particularly in countries where emission values are printed on cigarette packs. Consequently, these parameters may represent a useful framework with which to categorise and summarise toxicant emissions from cigarettes.

3.1.3. Cigarette smoke toxicant emissions as ratio to nicotine

Consistent with this possibility, the use of ratios of toxicants to nicotine has been proposed previously as a regulatory framework with which to compare and regulate cigarettes (Burns et al., 2008). This approach was proposed by TobReg, who selected the

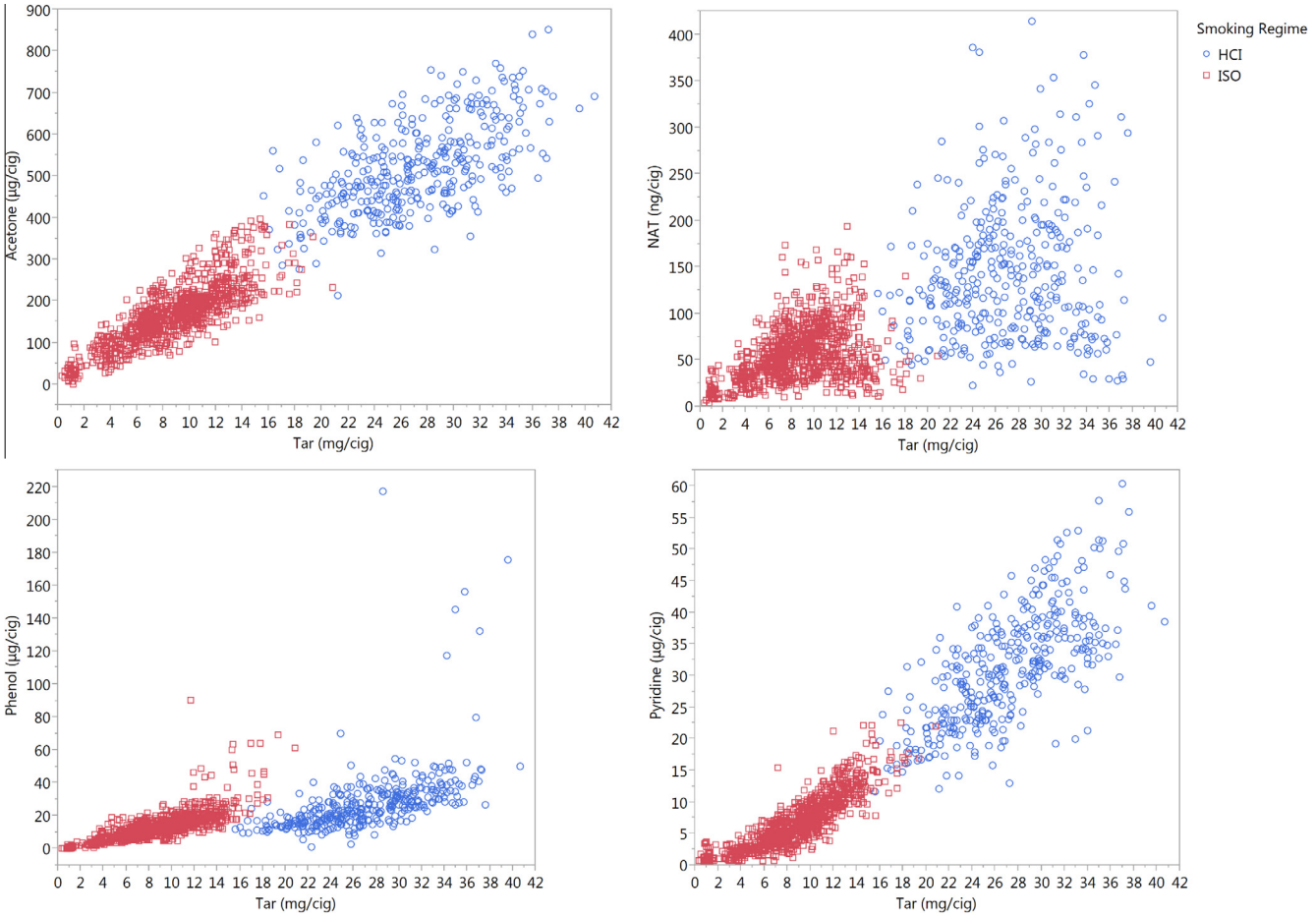

Fig. 2. Graphical representations of toxicant distributions and their relationships to tar for the ISO and HCI machine smoking regimes.

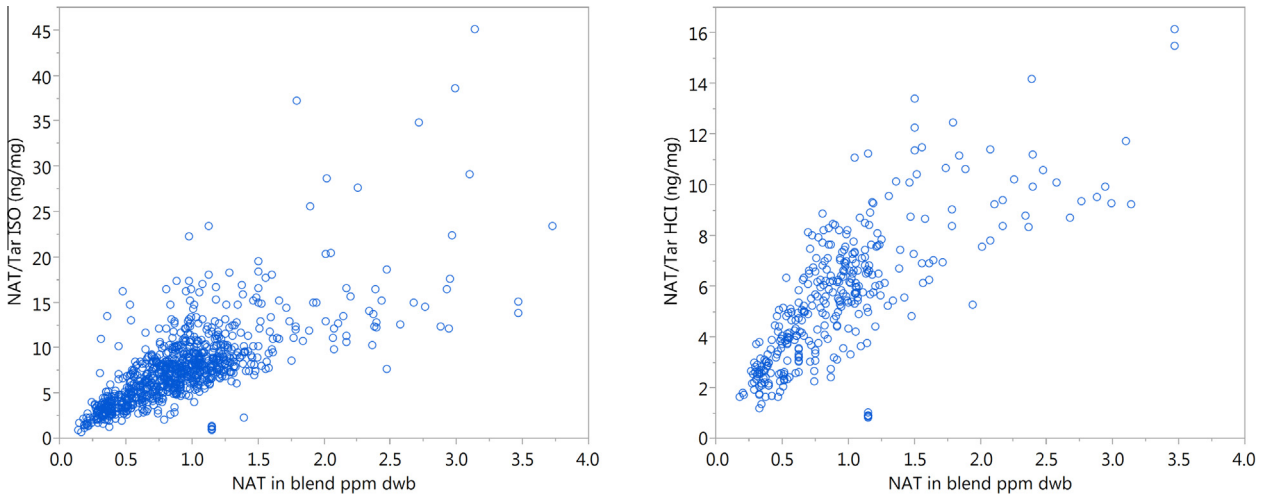

Fig. 3. NAT in blend with respect to NAT/tar measured using the ISO regime (left) and the HCI regime (right).

HCI regime as the basis of measurement and regulation. We examined this approach in the current study, as well as performing exploratory analysis using ISO data of ratios to nicotine. We used NAT as a test case.

The distributions of the two sets of ratios are shown in Fig. 4. Although the distributions appeared to look different, with some higher values for the ISO ratios that emphasised the skewness, the boxplot showed that the majority of the ratios were concentrated in the same region (approximately 10–180 ng/mg).

The visual similarities in distributions were translated into percentiles (Table 3). The percentiles for ratios of NAT to nicotine appeared to sit around the same region of values. As expected, differences between the HCI and ISO datasets appeared to be more obvious towards the edges. ISO estimates were based on

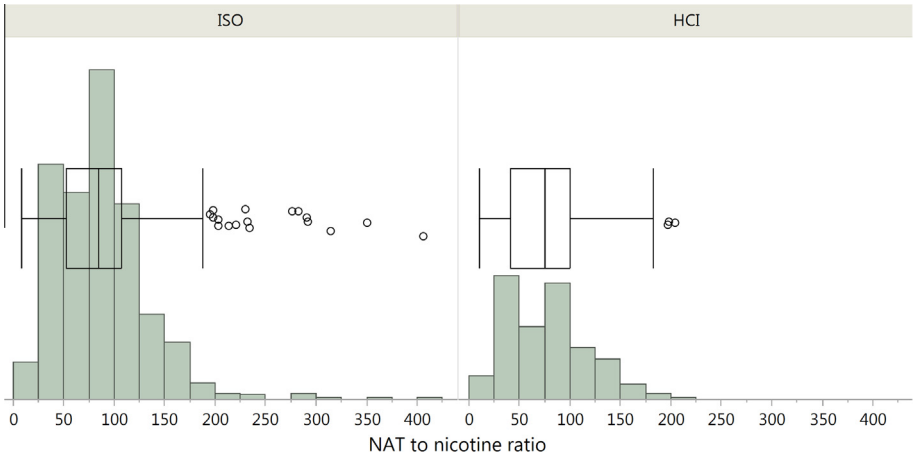

Fig. 4. Ratios of NAT to nicotine under ISO regime (left) and HCI regime (right) puffing conditions. Where the histogram's *y*-axis represents the number of products which, divided by the total number of products, can be seen as an empirical distribution of this sample. The boxplot is formed by the median (central line); the box is the interquartile range, whiskers are situated at 1.5 times from quartiles; observations outside the whiskers (right side points) are often interpreted as extreme values.

Table 3
Percentiles for NAT to nicotine ratios measured using ISO and HCI regimes.

Percentile	NAT to nicotine ratio	
	ISO	HCI
5th	28	24
10th	34	30
25th	53	41
40th	74	60
50th	84	76
60th	93	84
75th	107	101
90th	140	133
95th	161	149
99th	230	188

approximately three times more data than HCI estimates, which should be considered when assessing the estimates.

Percentiles based on ratios to nicotine are a good approximation to one parameter quantile regression estimates (Fig. 5). However, percentiles based on ratios do not adjust to specific data features independently of the density across the data space and therefore values for lower toxicant yield products will be less reliable, given that they will be more sensitive to errors in nicotine and toxicants. Comparisons of percentile ratios calculated using tar and nicotine showed a good level of comparability, but there were some discrepancies, especially in the range 6–12 mg tar (Fig. 5).

Observations in red are above the 90th percentile, calculated based on NAT to nicotine ratios. The line on the graph for NAT to nicotine (right) represents the 90th quantile for nicotine, with only nicotine as explanatory variable.

Although ratios of toxicants to tar or nicotine are robust approaches for representing levels of toxicants in cigarette smoke, they are very sensitive to small changes at low yields. Additionally, the ratio representation is not transparent for indicating the actual toxicant level, with implicit interpretation issues determining whether the product's toxicant yield is low or the nicotine/tar content is comparatively high.

3.2. Regression approach

3.2.1. Simple linear regression using tar as explanatory variable

In order to accommodate the sensitivity of toxicant yields to the overall levels of smoke emissions from cigarettes (as represented by tar), simple regression analysis with tar an explanatory variable

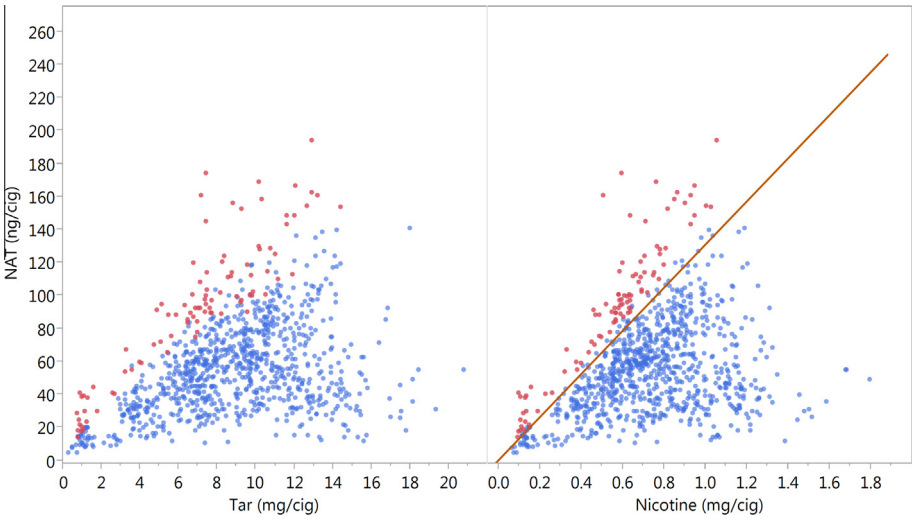

Fig. 5. NAT with respect to tar and nicotine.

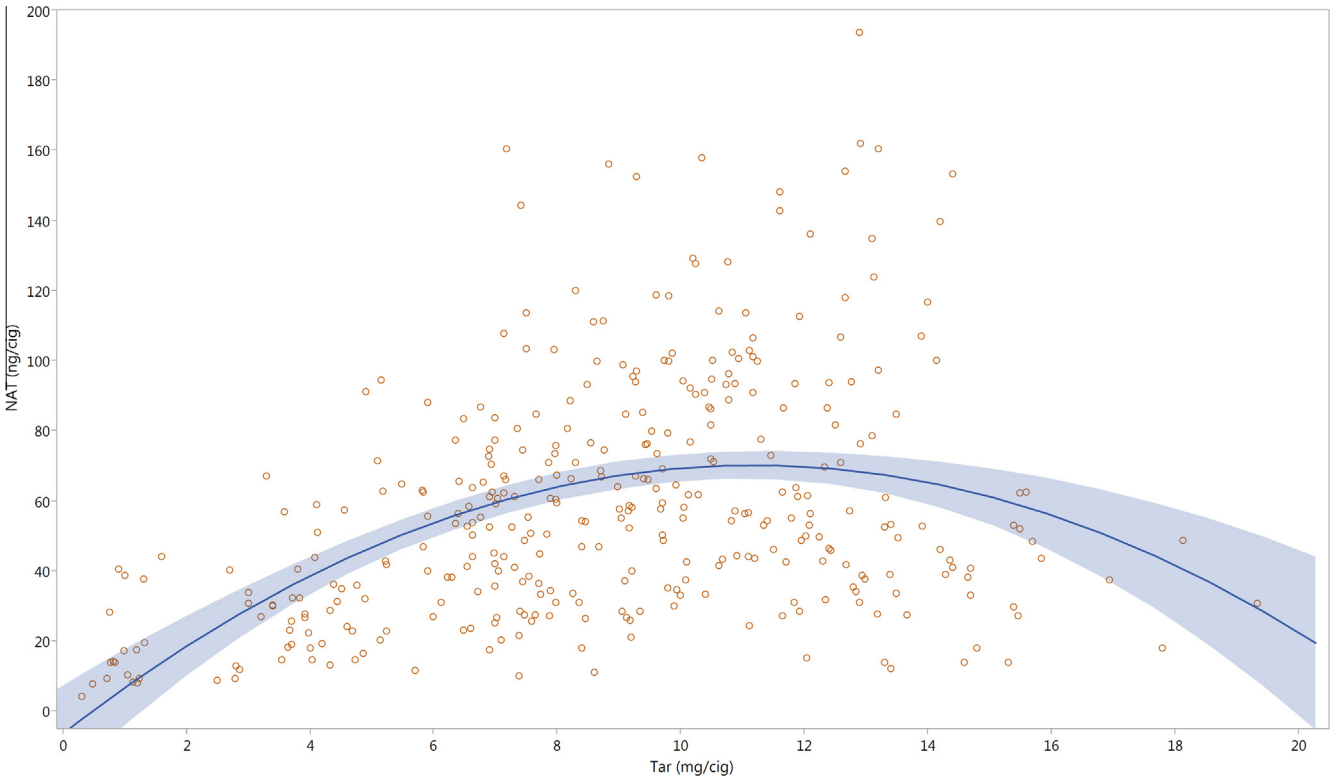

Fig. 6. Quadratic simple regression of NAT (measured using the ISO regime) with tar and its prediction confidence interval.

was explored (Fig. 6). The best polynomial fit of the regression between ISO NAT and tar was a quadratic fit, which is displayed in Fig. 6 with its 95% predictive confidence interval.

The quadratic approach provided the best fit to the measured data (although still with a very poor R-square = 0.24). The model appears to overfit the data, because the mean value falls after reaching a peak mean value for the population around the 11 mg tar yield. Alternatively, there may be a real change in cigarette blend character and reduction in NAT emissions for higher tar yield cigarettes. Although this approach allows calculation of predictions of the mean values at nominal tar references, the dispersion of data along the regression line broadens as tar increases up to approximately 8 mg tar, from which point onwards it becomes stable.

The most significant weakness of this approach is that it is based on a mean value, with much of the measured data containing significant levels of unexplained variation; a single reference value does not describe the diversity or distribution of the data yield around the regression line. Linear regression is also very sensitive to extreme values, which could provide unstable estimates if new data were added to the database.

3.2.2. Unconstrained quantile regression

In order to set up descriptive models for toxicant yields with respect to tar, we followed an empirical approach. The modelling process and assessment approach was identical for both regimes; here we discuss only the ISO regime to avoid repetition. As the toxicants measured with the ISO regime displayed different types of associations with respect to ISO tar (revealed as different shapes in Figs. 7a–d), we employed polynomial models, although different order polynomials were examined to accommodate different responses to tar. For example, phenol and pyridine might need second or third polynomial models to fit the data appropriately, whereas for acetone, a first order model might be sufficient. In addition, in unconstrained quantile regression, models are fitted independently from each other with the premise that different models might be more suitable for representing different quantiles. However, this flexibility for choosing different models introduces unnecessary subjectivity for establishing references. Given that the objective is to estimate quantile boundaries rather than act as explanatory models, we accepted a common polynomial model to be suitable across all quantiles for each toxicant. The model should be able to fit the data appropriately across quantiles without overfitting the data. Data overfitting leads to undesirable sensitivity of estimates to small data changes.

Based on growth curves (SAS Institute Inc., 2011), we commenced assessing a model in function of tar where the most common linear data shapes were considered, referred hereafter as "the full model". This full model contained variations up to the third order of tar. Where the inverse function of tar was able to represent initial down trends, first order for straight lines and the other parameters would indicate different curvatures. Initial empirical modelling suggested that second or third order variations of the full model were suitable candidates for all endpoints. An important assumption introduced in these models was that the toxicant to tar intercept was 0. In practical terms, we assumed that for tar equal to 0 (non-smoked products) there was no smoke and, therefore, the levels of smoke toxicants were also equal to 0. In modelling, this assumption implies that all models cross through the origin and thus will have an impact on parameter estimates.

We examined the impact of the three polynomial models on reference estimates. Table 4 displays the median bootstrap estimates of the three regression models for NAT at 19 ISO tar yield reference levels and their respective coefficients of variation. Only two observations were above the upper tar value (19 mg). Although model elections had a clear impact on estimates, the estimates of the medians appeared to be comparable for cigarettes with tar reference levels that had a high density of data in the data space of products (3–15 mg tar), whereas differences were more evident for cigarettes with tar yields at the upper or lower end of the reference range. This effect was especially noticeable for

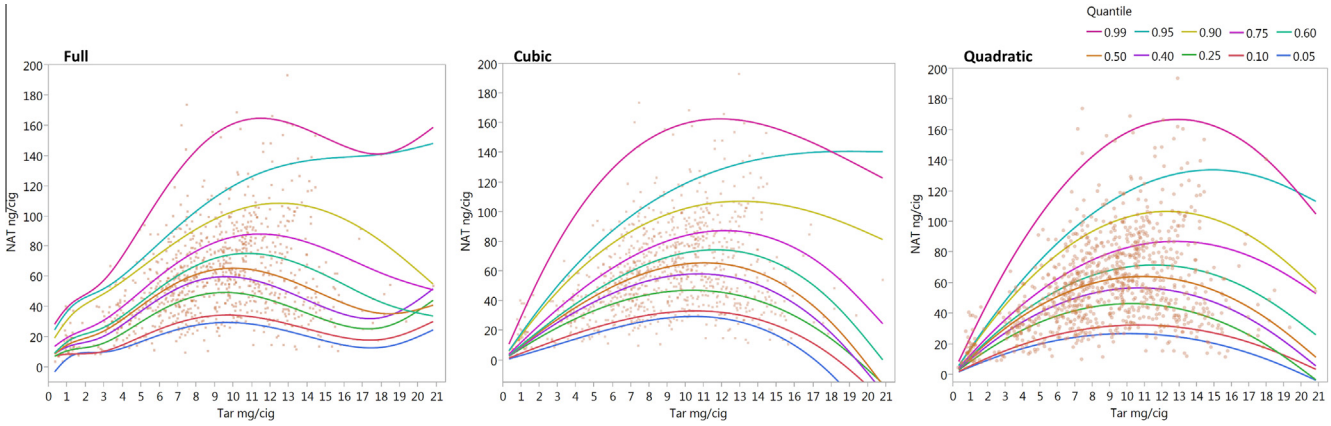

Fig. 7A. Unconstrained quantile regressions fitted for ISO NAT with respect to ISO tar, for full, cubic and quadratic models (respectively, from left to right).

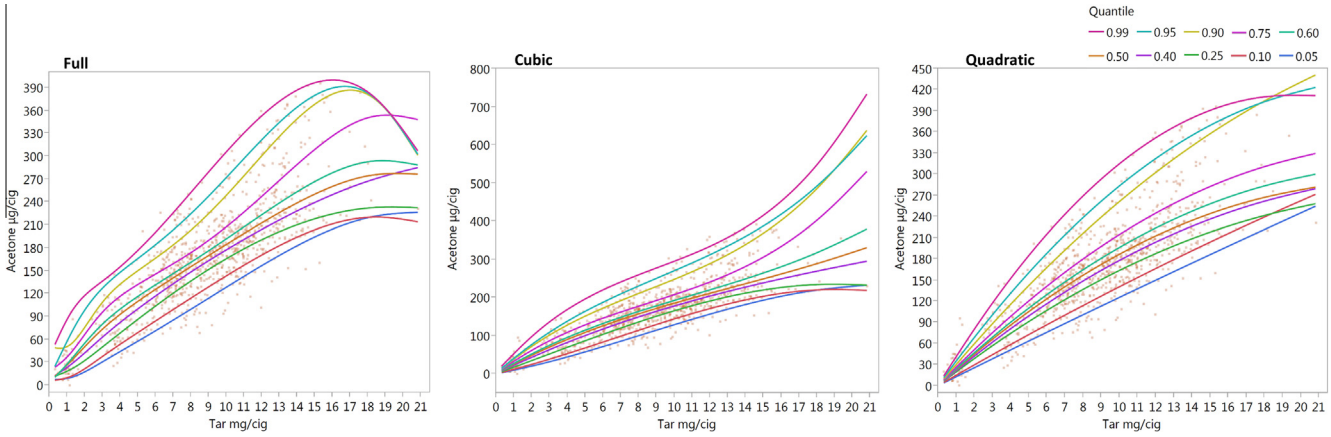

Fig. 7B. Unconstrained quantile regressions fitted for ISO acetone with respect to ISO tar, for full, cubic and quadratic models (respectively, from left to right).

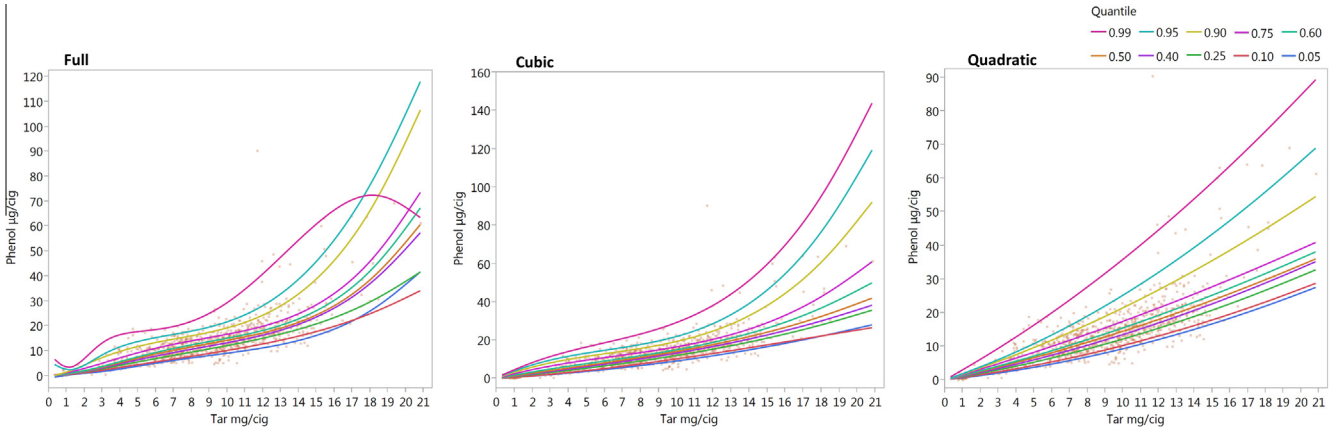

Fig. 7C. Unconstrained quantile regressions fitted for ISO phenol with respect to ISO tar, for full, cubic and quadratic models (respectively, from left to right).

cigarettes with tar < 3 mg and >15 mg. Low tar estimates responded to small changes in the lower bound of the bootstrap samples to comply with the intercept assumption. In addition, areas around tar references with sparse data yielded higher variability of estimates across models (Table 4). Coefficients of variation increased in areas with decreasing data density. The variation in estimates was especially high for cigarettes with >15 mg tar, which made the results unreliable. These observations showed that the dataset at the high tar areas was insufficient to clearly define the toxicant tar responses. Consequently, the decision was made to exclude data >15 mg tar from subsequent calculations. The variation at <3 mg tar was deemed to be acceptable, although estimates should be used with caution, especially below the 10th and above the 90th percentiles, where several coefficients of variation (CVs) were above 10%.

The three models for all four endpoints are graphically represented in Fig. 7A–D. Although in empirical modelling the simplest model is generally preferable, here the quadratic model did not fit

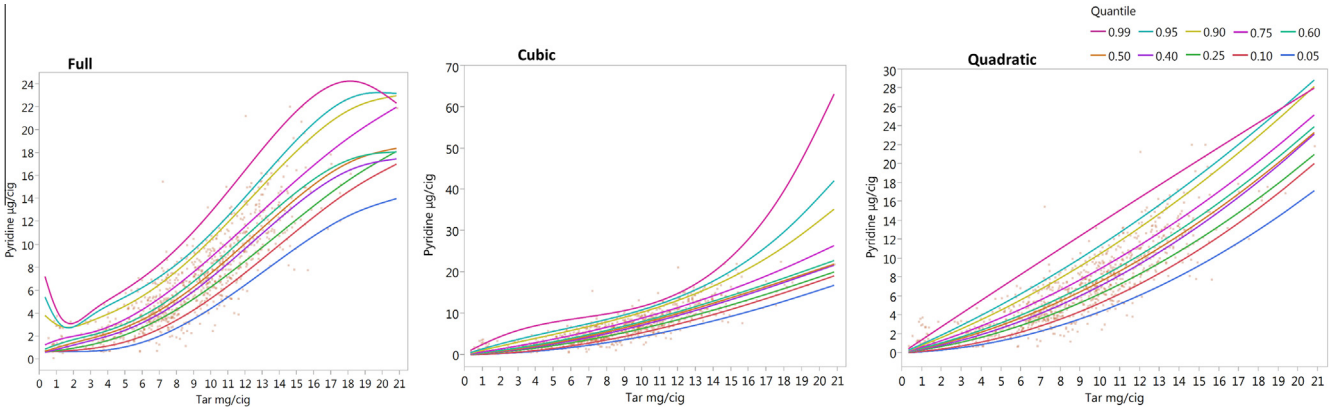

Fig. 7D. Unconstrained quantile regressions fitted for ISO pyridine with respect to ISO tar, for full, cubic and quadratic models (respectively, from left to right).

most of the local features of the data and the full model was over-parameterised. The quantiles obtained by the quadratic model appeared to cross the data almost in straight lines for some toxicants, which appeared to disagree with some of the patterns observed in Fig. 7, whilst the full model appeared to be too sensitive to data scarcity. This is shown in Fig. 7 (left column) with data attempting to fit every local feature, which may lead to lack of generality of model estimates. This effect was especially noticeable for NAT, for which oscillating patterns appeared throughout the tar bands, whereas for the other toxicants this occurred at the higher tar bands, where the full model "wave" was attempting to adjust to the low number of observations. Additionally, for the full model there were decreasing patterns for the 0.90 and 0.95 quantiles for cigarettes with low tar yields. The fits of the cubic and quadratic models were comparable within the well-characterised range of tar yields (1–15 mg tar). However, the cubic model can adjust to local features and thus can fit a broader variety of shapes beyond the four exemplar toxicants assessed here. The CVs in Table 4 were also considered to be comparable with, or below the full model. Most coefficients were under 10% variation, with a few exceptions at 1 mg and 2 mg tar. Hence, for our purpose, we considered the cubic model to provide adequate fit in well characterised ranges of toxicants across quantiles.

Crossing of quantile lines was observed for all four analytes (Figs. 7a–d). Although this crossing effect happened at tar yields >15 mg for all toxicants in all models, it is not guaranteed that it will not happen within the lower range of tar yields for any other analyte, and it would invalidate interpretation of the estimates.

The NAT data exemplifies other undesirable features of unconstrained quantile regression for this specific application. A lack of data at high tar levels drove down estimates towards high leverage points near the highest tar values. NAT quantiles reach an inflexion point at 11–12 mg tar. In some cases, as explained in the discussion section, there could be reasons to believe that cigarettes with high tar levels would yield lower levels of NAT than cigarettes with lower tar levels. However, a possible simple explanation is that it is the result of scarcity of data.

3.2.3. Non-crossing quantile regression

Linear constraints were introduced to ensure non-crossing of quantile regression lines and monotonicity with respect to tar. However, the constraint that prevents crossing of quantiles clashes with the assumption that all quantile curves have a common origin at tar = 0. Therefore, the non-crossing quantile model included an intercept:

$$\text{Toxicant} = \beta_0 + \beta_3\text{tar} + \beta_5\text{tar}^2 + \beta_6\text{tar}^3 \quad (5)$$

The inclusion of an intercept could present an interpretation issue at tar = 0; however, possible reasons for inclusion of an intercept are explained in the discussion. Taking a conservative approach, in our model tar = 0 implies that cigarettes have not been smoked and therefore, is discrete at tar = 0. That is, the relationship between the toxicants and tar is defined as:

$$\begin{cases} \text{Toxicant} = 0 \text{for tar} = 0 \\ \text{Toxicant} = \beta_0 + \beta_3\text{tar} + \beta_5\text{tar}^2 + \beta_6\text{tar}^3 \text{for tar}(0, +\infty) \end{cases} \quad (6)$$

3.2.3.1. Toxicants measured with the ISO regime. The results of fitting this model to the ISO data are displayed in Fig. 8. The monotonicity constraint is clearly exemplified by NAT, for which quantiles become constant after reaching a maximum estimate. The non-crossing relationship of quantiles is achieved within the data space; the highest quantiles for phenol cross outside the data boundaries and other quantiles suggest interception may occur outside the fitted data range.

Estimates corresponding to the median of the 10,000 iterations per tar reference value and bootstrap confidence intervals for NAT are reported in Table 5A. For comparison between estimates, the values for NAT from the unconstrained quantile regression are shown in Table 5B.

Non-crossing model estimates for NAT yields from 1 to 2 mg tar products were higher than those estimated from the crossing model because of the introduction of the intercept. However, the values were comparable for other tar level estimates. The non-crossing model appeared to yield consistently lower estimates for lower tar values (left side of the table) than the crossing model. This difference is the result of the combined effect of higher initial values for cigarettes with 1–2 mg tar values, and monotonicity adjustments on the right side of the table. Differences towards the edges were more pronounced, with up to a 21 ng difference for three estimates: two values at the 99th percentile (ISO tar 4 mg/cig and 5 mg/cig) and one at the 5th percentile of the lowest tar level. The average difference at the edges was 6.1 ng, while in the central region this decreased to 2.2 ng, highlighting the robustness of estimates for higher data density in the central area.

3.2.3.2. Toxicants measured with the HCI regime. The same modelling process and assumptions were also used for the HCI regime data and a cubic polynomial model was also chosen as the standard model for the HCI analytes. However, the data spaces for the ISO and HCI regimes were very different. The HCI tar data range was 14–44 mg/cig. Therefore, to take into account increased variation in estimates at the lower and upper limits of data availability and to create a comparative table, we set the reference value range

Table 4
Percentile estimates for NAT (ng/cig) under ISO smoking conditions for three models (full, cubic, quadratic). Estimates are the median (% CV) of 10,000 bootstrap iterations at each tar level. A full tabulation may be found in Supplemental data (Table App2).

Tar (mg/cig)	Percentile and model														
	5th			25th			50th			75th			95th		
	Full	Cubic	Quad	Full	Cubic	Quad	Full	Cubic	Quad	Full	Cubic	Quad	Full	Cubic	Quad
1	8 (12)	4 (16)	5 (5)	12 (13)	8 (8)	9 (4)	16 (9)	10 (8)	11 (3)	21 (17)	12 (7)	13 (4)	39 (11)	18 (10)	18 (5)
3	10 (10)	11 (10)	14 (4)	16 (13)	22 (5)	23 (3)	24 (8)	28 (5)	30 (2)	33 (9)	35 (5)	37 (4)	51 (8)	50 (7)	49 (4)
5	17 (5)	18 (5)	20 (4)	30 (6)	33 (3)	34 (3)	40 (3)	43 (3)	45 (2)	50 (4)	54 (3)	55 (3)	71 (5)	77 (5)	75 (3)
7	25 (4)	25 (3)	24 (3)	43 (3)	42 (3)	42 (3)	56 (2)	55 (2)	56 (2)	69 (3)	69 (2)	70 (3)	93 (3)	97 (3)	97 (3)
9	29 (4)	29 (4)	27 (3)	49 (3)	47 (3)	46 (3)	65 (2)	63 (2)	62 (2)	82 (3)	80 (2)	80 (2)	113 (3)	113 (3)	113 (2)
11	29 (6)	29 (5)	27 (4)	48 (3)	47 (4)	46 (3)	65 (3)	66 (3)	64 (2)	88 (3)	87 (3)	86 (2)	125 (4)	123 (4)	125 (3)
13	24 (11)	26 (6)	25 (8)	40 (4)	44 (4)	43 (4)	58 (5)	63 (4)	62 (3)	86 (5)	87 (4)	87 (4)	131 (6)	127 (5)	130 (5)
15	18 (16)	19 (14)	20 (15)	31 (6)	35 (8)	36 (6)	48 (8)	54 (7)	55 (6)	78 (8)	82 (7)	84 (6)	129 (9)	131 (8)	132 (8)
17	14 (52)	4 (119)	14 (30)	26 (9)	23 (27)	26 (12)	37 (12)	37 (17)	44 (10)	64 (15)	69 (14)	78 (10)	130 (17)	137 (14)	127 (12)
19	17 (162)	-16 (-68)	6 (93)	29 (19)	6 (518)	12 (37)	36 (24)	12 (83)	29 (22)	50 (30)	49 (31)	67 (17)	126 (37)	140 (25)	118 (18)

CV, coefficients of variation; A negative mean at tar 19 makes the CV to be also negative which doesn't have an interpretation as dispersion measurement as well as having a negative median value.
Tar1, Tar2 = ISO tar 1 mg/cig, 2 mg/cig, etc.

at 17–38 mg/cig, with 3 mg intervals. Given that this was a smaller dataset than ISO and it was defined over a wider range, the variability of estimates for HCI data was expected to be larger than for ISO data. Fig. 9 displays HCI NAT data assessed with a cubic polynomial model, with and without monotonicity and non-crossing constraints.

Even here, where there was a poor relationship between tar and NAT levels (spearman $R = 0.0734$), quantile regression models were able to fit the data. In the non-crossing model, horizontal lines suggest a lack of relationship between the variables for most of the quantiles, with slightly increasing patterns for the lower tar levels. However, the crossing model was again likely to provide underestimates for toxicants at high tar levels (Table App4).

3.3. Product comparisons with quantile regression estimates

As we have shown, understanding of the toxicant yields of current commercially available products can be enhanced by the use of tabulated percentiles at nominal tar values. Comparisons between products can be performed by approximation to the nominal value, or by linear interpolation to intermediate tar values.

Those quantile regression models we have used to calculate the reference tables can also be used to estimate percentiles for the actual tar yield of the product being assessed. For example, given an exemplar product A, with a 9.8 mg/cig ISO tar yield, the 5th, 10th, 25th, 40th, 50th, 60th, 75th, 90th, 95th and 99th percentiles for NAT, acetone, phenol and pyridine have been calculated (using non-crossing models) and are displayed in Table 6.

If the toxicant yields of product A were: NAT 102 ng/cig; acetone 156 μg/cig; phenol 10.8 μg/cig; and pyridine 4.3 μg/cig; then it would be >90th percentile for NAT, below the lower quartile for acetone and phenol, and between the 5th and 10th percentile of products at this tar level for pyridine. Therefore, this approach allows direct comparisons without the need for interpolations.

This approach considers one smoke toxicant at a time. As a more sophisticated approach, even when the relationship of toxicants is not necessarily linear, an average percentile across several toxicants could be informative as a reference. This builds on a previous concept by McAdam et al. (2012), which ranked products across a range of toxicant yields.

An average value could be further generalised by incorporation of weighting factors, for example, indicating different priority toxicants and/or biological relationships between toxicants. This approach is compatible with previous work on prioritisation of smoke toxicants using the Margin of Exposure model (Cunningham et al., 2011; Soeteman-Hernandez et al., 2013) or hazard indexes (Haussmann, 2012). The "overall percentile" for a product can be written as:

$$\text{Product percentile} = \sum w_i * \text{PercTox}_i \quad (7)$$

where i represents a toxicant and w_i the weighting factors $\sum w_i = 1$.

This approach could be similarly implemented for nicotine as the reference variable.

4. Discussion

Compared with traditional approaches for the calculation of reference values (Conaway, 2009), quantile regression methodology presents several advantages for the comparison of smoke toxicant yields from commercial cigarettes. Comparisons of commercial products can be interpreted intuitively by the assessment of percentiles instead of averages within a comprehensive framework. Another advantage is that non-parametric quantile regression does not assume an underlying distribution and therefore data transformation is not required (Marrie et al., 2009).

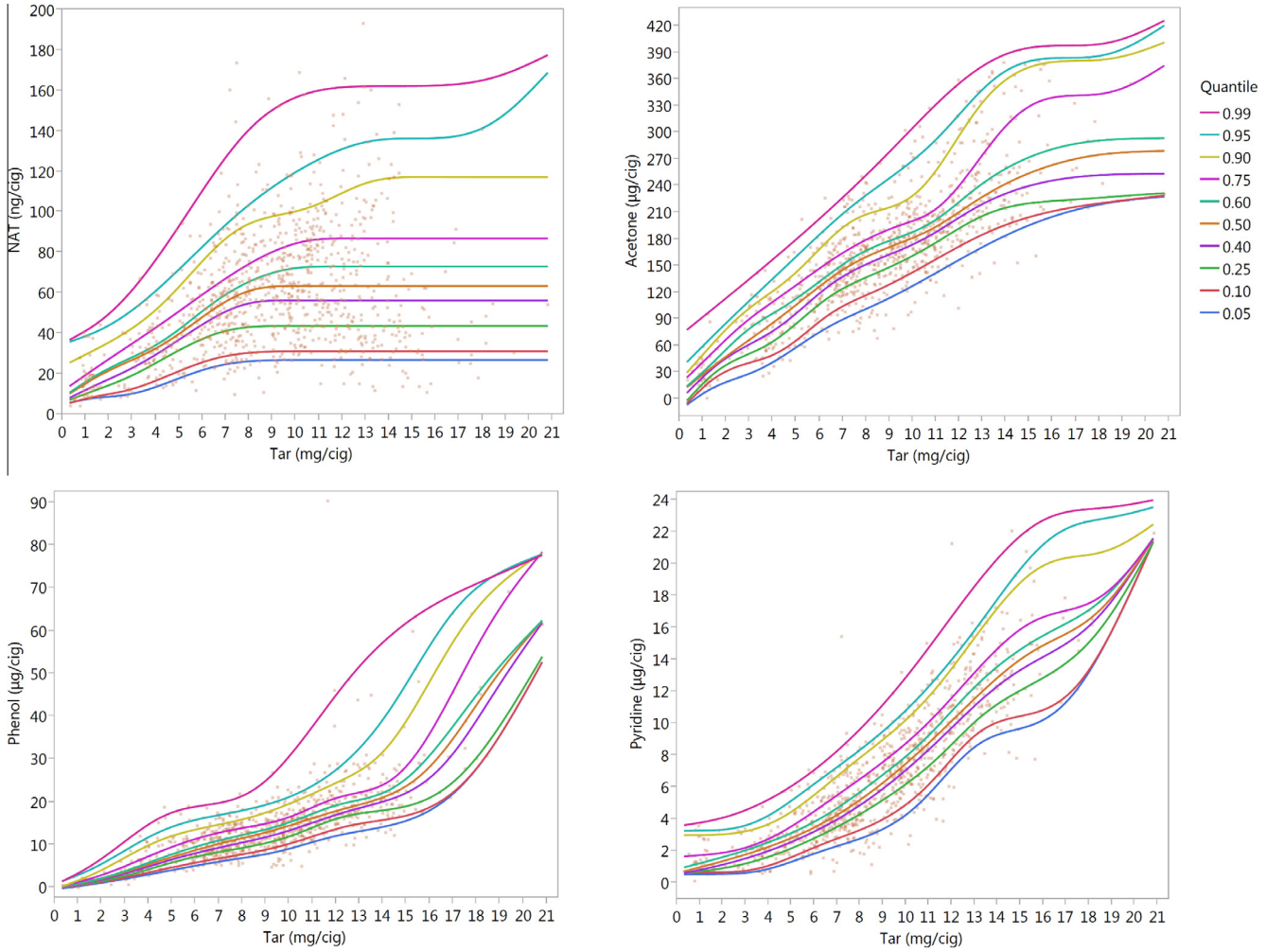

Fig. 8. ISO toxicants with respect to ISO tar non-crossing and monotonicity constrained quantiles fitted using a cubic regression model.

Table 5A
Percentile estimates for NAT, ng/cig under ISO smoking condition, including bootstrap confidence intervals using quantile regression with non-crossing and monotonicity constraints. A full tabulation is available in Supplementary data (Table App3A).

Median (CI)	ISO tar							
	1 mg/cig	3 mg/cig	5 mg/cig	7 mg/cig	9 mg/cig	11 mg/cig	13 mg/cig	15 mg/cig
99th	41 (37, 43)	62 (51, 65)	93 (78, 106)	127 (106, 152)	151 (123, 165)	160 (138, 176)	162 (152, 195)	162 (153, 206)
95th	39 (26, 41)	51 (43, 60)	71 (64, 82)	94 (90, 104)	112 (101, 117)	123 (110, 134)	134 (118, 152)	137 (119, 154)
75th	19 (17, 26)	35 (30, 39)	50 (45, 55)	68 (64, 73)	81 (76, 86)	87 (82, 91)	87 (83, 95)	87 (83, 96)
50th	15 (12, 17)	27 (23, 30)	39 (37, 42)	56 (53, 58)	63 (61, 65)	63 (61, 67)	63 (61, 67)	63 (61, 67)
25th	10 (8, 13)	18 (14, 23)	32 (27, 36)	41 (38, 44)	44 (41, 47)	44 (41, 47)	44 (41, 47)	44 (41, 47)
10th	8 (6, 10)	12 (10, 15)	21 (18, 25)	29 (26, 31)	31 (29, 33)	31 (30, 34)	31 (30, 34)	31 (30, 34)
5th	7 (5, 9)	11 (10, 13)	18 (16, 20)	25 (22, 26)	27 (24, 28)	27 (24, 29)	27 (24, 29)	27 (24, 29)

We found that although percentile estimates were dependent on which quantile regression model was chosen, subjectivity could be overcome by acceptance of a common model across percentiles and even across toxicants. For a chosen model, quantile regression provided robust reference estimates, although variability increased for estimates approaching the extremes of the data space. In general, a polynomial cubic model adjusted to local features of the data, while variability remained acceptable. Cigarettes with ISO tar yields of 1–15 mg/cig tar, which fall into the 5th to 95th percentiles, seem to be well characterised for toxicants measured under the ISO regime, using a criterion of CVs < 10%. There were some exceptions for low percentiles of tar values that approached

Table 5B
Percentile estimates for NAT, ng/cig under ISO smoking conditions, including bootstrap confidence intervals using unconstrained "crossing" quantile regression. A full tabulation is available in Supplementary data (Table App3B).

Median (CI)	ISO tar							
	1 mg/cig	3 mg/cig	5 mg/cig	7 mg/cig	9 mg/cig	11 mg/cig	13 mg/cig	15 mg/cig
99th	30 (22, 29)	78 (62, 98)	114 (94, 137)	140 (113, 159)	154 (126, 168)	160 (138, 171)	161 (145, 195)	156 (134, 231)
95th	18 (16, 23)	50 (47, 60)	77 (73, 85)	97 (92, 103)	113 (106, 119)	123 (114, 130)	127 (117, 139)	131 (112, 148)
75th	12 (10, 14)	35 (31, 39)	54 (50, 58)	69 (66, 73)	80 (77, 84)	87 (83, 92)	87 (81, 94)	82 (72, 93)
50th	10 (8, 11)	28 (25, 31)	43 (41, 46)	55 (53, 57)	63 (60, 66)	66 (62, 70)	63 (59, 69)	54 (48, 63)
25th	8 (7, 9)	22 (20, 25)	33 (32, 36)	42 (40, 44)	47 (43, 49)	47 (44, 50)	44 (40, 47)	35 (31, 40)
10th	5 (3, 6)	14 (11, 16)	22 (20, 25)	29 (26, 31)	33 (30, 36)	34 (31, 37)	31 (28, 35)	24 (17, 31)
5th	4 (3, 5)	11 (9, 14)	18 (17, 20)	25 (23, 26)	29 (26, 30)	29 (26, 32)	26 (23, 29)	19 (15, 26)

CI, confidence intervals.

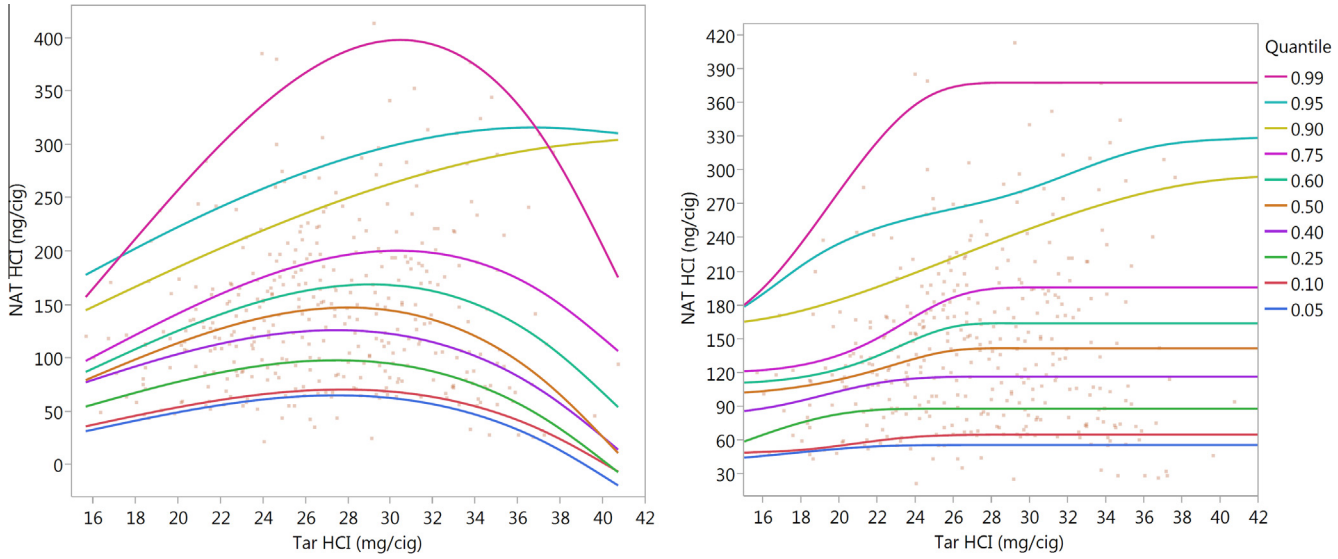

Fig. 9. Unconstrained quantile regressions for NAT and tar measured under the HCI regime (left) and monotonicity and non-crossing quantile regressions for NAT HCI versus tar HCI (right).

the lower bound. For HCI measurements, toxicant estimates seemed to be robust when tar references values of 18–36 mg/cig were used (data not shown).

Undesirable features of quantile regression, such as lines crossing, can be avoided by the introduction of simple linear constraints. High leverage points could have a high impact on percentile estimates. Inclusion of new data points to fill scant areas along the *x*-axis is likely to reduce this effect. For some toxicants, leverage also produces a downwards trend for the less populated data region at high levels of tar. In the absence of an explanation for these downward trends they can be assumed to be artefacts driven by local data leverage, and therefore, increasing monotonicity constraints can be introduced.

Reference tables can be used to compare products by toxicant emission levels. Careful consideration should, however, be given to the way in which these tables are used to assess smoke toxicant yields of cigarettes. The tables provide guidance on whether the specific toxicant yield for a product is within the expected reference values for that toxicant, but they do not provide information about the combined effect of several toxicant yields. Prioritisation of toxicants with respect to health risk, in conjunction with percentile estimates, could be used to develop an overall toxicant score to simplify and add objectivity to comparisons of commercially available cigarettes.

Table 6
ISO toxicants percentiles for a product with a 9.8 mg/cig ISO tar yield.

Percentile	Analyte (9.8 mg/cig tar)			
	NAT (ng/cig)	Acetone (μg/cig)	Phenol (μg/cig)	Pyridine (μg/cig)
5th	27.0	124.0	8.7	4.1
10th	31.0	139.0	9.8	4.7
25th	44.0	158.0	11.5	6.0
40th	56.0	171.0	12.9	6.8
50th	63.0	179.0	14.0	7.2
60th	72.0	186.0	14.8	7.7
75th	84.0	198.0	16.0	8.5
90th	100.0	225.0	19.0	10.0
95th	118.0	264.0	20.7	10.6
99th	156.0	300.0	29.4	12.6

This empirical assessment of quantile regression for estimating reference percentiles of smoke toxicants for different endpoints was only based on non-parametric quantile regression (Koenker and Basset, 1978). This approach was chosen because it was able to fit smoke toxicant data adequately and implementation was

readily available with mainstream software. Approaches based on kernel density distribution curves (Gannoun et al., 2002) and with parametric assumptions of the underlying distribution, such as the least mean square method (Cole, 1988; Cole and Green, 1992) and using maximum likelihood (Noufaily and Jones, 2013), are some examples of the many possible variants for alternative quantile regression analyses. Wei et al. compared some of these methods (Wei et al., 2006), while avoidance of regression crossing in quantile regression constitutes a field of research in its own right (Bondell et al., 2010; Liu and Wu, 2011).

Currently, data entries are considered to be independent, but this may not strictly be the case. For example, the same brand variant could be analysed in consecutive years to create a set of non-independent observations. Longitudinal methodology (Koenker, 2004; Wei et al., 2006), which accounts for time-related dependences, may help to reduce the width of confidence intervals by providing more accurate estimates of percentiles. At this point, the data may become very complex because despite some of these data entries belonging to the same brand variant, specification changes over time mean that subsequent samples are not strictly equivalent, e.g. due to changes in blend composition. It then becomes very difficult to determine when observations within the same brand variant should be considered dependent or independent from each other.

5. Conclusions

In this paper, we examined common approaches for analysing and representing toxicant yields in cigarette smoke and we introduced quantile regression as a new approach with which to understand the diversity of toxicant yields from cigarettes. To assess these approaches, we used the most complete database of smoke and tobacco blend constituents reported to date.

Percentiles were shown to be an intuitive and robust way to summarise the range of tobacco blend toxicant precursors, and the ratios of smoke toxicants to nicotine or tar. However, there were drawbacks, as ratios do not adjust to specific data features; in principle they are more sensitive to potential errors in toxicants or reference analyte measures, and they also normalise products with high levels of both toxicant and reference analyte (nicotine or tar), whilst tending to de-normalise products with low levels of the reference analyte. Classical regression approaches (both mean and median based) presented limitations for representing the range of diversity found in the data for smoke constituents, and means were sensitive to extreme values.

As an extension of the median approach, we showed that quantile regression, either constrained or unconstrained, can be used to effectively represent data distribution of toxicants measured in either ISO or HCI regimes. Some of the advantages of this method are that it uses the complete dataset to estimate percentiles at nominal tar values rather than intervals, it does not require data transformation, and it provides robust, transparent and intuitive percentile estimates in relation to any desired reference value within the data space. However, it is also associated with drawbacks, such as its sensitivity to the model used, and the occurrence of crossing curves with unconstrained regressions. However, overall, quantile regression presents some significant advantages with respect to current approaches for representing smoke toxicant yields from contemporary tobacco products.

Role of the funding source

This study was funded by British American Tobacco.

Conflict of interest

All authors are employees of British American Tobacco.

Transparency Document

The Transparency document associated with this article can be found in the online version.

Acknowledgements

The authors thank Vito M. R. Muggeo for his assistance with the R package quantregGrowth. Editorial assistance with formatting and proofing of this manuscript was provided by JEM Comms Ltd. and funded by British American Tobacco. We would also like to thank the three reviewers whose suggestions have helped to add clarity to this publication.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.yrtph.2015.05.023.

References

AOAC International, 2002. Guidelines for single laboratory validation of chemical methods for dietary supplements and botanicals. <http://www.aoac.org/imis15_prod/AOAC_Docs/StandardsDevelopment/SLV_Guidelines_Dietary_Supplements.pdf> (Accessed: December 2014).

Australian Government Department of Health and Ageing, 2002. <http://www.health.gov.au/internet/main/publishing.nsf/Content/tobacco-emis> (Accessed: March 2011).

Baker, D.L., Krol, E.S., Jacobsen, N., Liebler, D.C., 1999. Reactions of beta-carotene with cigarette smoke oxidants. Identification of carotenoid oxidation products and evaluation of the prooxidant/antioxidant effect. Chem. Res. Toxicol. 12, 535–543.

Baker, R., 2002. The development and significance of standards for smoking-machine methodology. Beitr. Tabakforsch Int. 20, 23–41.

Baker, R.R., 2006. The generation of formaldehyde in cigarettes–overview and recent experiments. Food Chem. Toxicol. 44, 1799–1822.

Benatar, M., Wuu, J., Peng, L., 2009. Reference data for commonly used sensory and motor nerve conduction studies. Muscle Nerve 40, 772–794.

Bodnar, J.A., Morgan, W.T., Murphy, P.A., Odgen, M.W., 2012. Mainstream smoke chemistry analysis of samples from the 2009 US cigarette market. Regul. Toxicol. Pharmacol. 64, 35–42.

Bondell, H.D., Reich, B.J., Wang, H., 2010. Noncrossing quantile regression curve estimation. Biometrika 97, 825–838.

Bouyé, E., Salmon, M., 2009. Dynamic copula quantile regressions and tail area dynamic dependence in Forex markets. Eur. J. Financ. 15, 721–750.

Burns, D.M., Dybing, E., Gray, N., Hecht, S., Anderson, C., Sanner, T., O'Connor, R., Djordjevic, M., Dresler, C., Hainaut, P., Jarvis, M., Opperhuizen, A., Straif, K., 2008. Mandated lowering of toxicants in cigarette smoke: a description of the World Health Organization TobReg proposal. Tob. Control 17, 132–141.

Cade, B.S., 2011. Estimating equivalence with quantile regression. Ecol. Appl. 21, 281–289.

Carpenter, J., Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. 19, 1141–1164.

Cole, T., 1988. Fitting smoothed centile curves to reference data. J. R. Stat. Soc. A 151 (part 3), 385–418.

Cole, T.J., Green, P.J., 1992. Smoothing reference centile curves: the LMS method and penalized likelihood. Stat. Med. 11, 1305–1319.

Conaway, M., 2009. Reference data and quantile regression. Muscle Nerve 40, 751–752.

Counts, M.E., Morton, M.J., Laffoon, S.W., Cox, R.H., Lipowicz, P.J., 2005. Smoke composition and predicting relationships for international commercial cigarettes smoked with three machine-smoking conditions. Regul. Toxicol. Pharmacol. 41, 185–227.

Cunningham, F.H., Fiebelkorn, S., Johnson, M., Meredith, C., 2011. A novel application of the Margin of exposure approach: segregation of tobacco smoke toxicants. Food Chem. Toxicol. 49, 2921–2933.

Fowles, J., Dybing, E., 2003. Application of toxicological risk assessment principles to the chemical constituents of cigarette smoke. Tob. Control 12, 424–430.

Gannoun, A., Girard, S., Guinot, C., Saracco, J., 2002. Reference curves based on non-parametric quantile regression. Stat. Med. 21, 3119–3135.

Gregg, E., Hill, C., Hollywood, M., Kearney, M., McLaughlin, D., McAdam, K., Williams, M., Purkis, S., 2004. The UK smoke constituents testing study. Summary of results and comparison with other studies. Beitr. Tabakforsch Int. 21, 117–138.

Hatsukami, D.K., Hecht, S.S., Hennrikus, D.J., Joseph, A.M., Pentel, P.R., 2003. Biomarkers of tobacco exposure or harm: application to clinical and

epidemiological studies. 25–26 October 2001, Minneapolis. Minnesota. Nicotine Tob. Res. 5, 387–396.
Haussmann, H.J., 2012. Use of hazard indices for a theoretical evaluation of cigarette smoke composition. Chem. Res. Toxicol. 25, 794–810.
Health Canada, 1999. Official method: determination of nitrosamines in mainstream smoke. No. T-111. Health Canada, Ottawa, Canada.
Hecht, S.S., 2012. Research opportunities related to establishing standards for tobacco products under the Family Smoking Prevention and Tobacco Control Act. Nicotine Tob. Res. 14, 18–28.
Hecht, S.S., Yuan, J.M., Hatsukami, D., 2010. Applying tobacco carcinogen and toxicant biomarkers in product regulation and cancer prevention. Chem. Res. Toxicol. 23, 1001–1008.
Hyodo, T., Maruta, Y., Itaya, H., Mikita, A., Kodera, T., Meger, M., 2007. Evaluation of functional relationships for predicting mainstream smoke constituent machine yields for conventional cigarettes from the Japanese market. Regul. Toxicol. Pharmacol. 48, 194–224.
International Organization for Standardization, 1994. ISO 5725–2:1994. Accuracy (trueness and precision) of measurements methods and results – part 2. Basic method for determination of repeatability and reproducibility of a standard measurement method. International Organization for Standardization, Geneva, Switzerland.
International Organization for Standardization, 2000. Routine analytical cigarette-smoking machine – definitions and standard conditions ISO 3308. International Organization for Standardization, Geneva, Switzerland.
Kim, M.O., Yang, Y., 2011. Semiparametric approach to a random effects quantile regression model. J. Am. Stat. Assoc. 106, 1405–1417.
Koenker, R., 2004. Quantile regression for longitudinal data. J. Multivariate Anal. 91, 74–89.
Koenker, R., 2013. quantreg: Quantile Regression. R package version 5.05. <http://crantastic.org/packages/quantreg> (Accessed: December 2014).
Koenker, R., Basset, G., 1978. Regression quantiles. Econometrica 46, 33–50.
Kroenger, R., 2005. Quantile regression. Econometric society monographs. Cambridge University Press, New York, USA.
Kuczmarski, R.J., Ogden, C.L., Guo, S.S., Grummer-Strawn, L.M., Flegal, K.M., Mei, Z., Wei, R., Curtin, L.R., Roche, A.F., Johnson, C.L., 2002. 2000 CDC Growth Charts for the United States: methods and development. Vital Health Stat. 11, 1–190.
Liu, Y., Wu, Y., 2011. Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. J. Nonparametric Stat. 23, 415–437.
Marrie, R.A., Dawson, N.V., Garland, A., 2009. Quantile regression and restricted cubic splines are useful for exploring relationships between continuous variables. J. Clin. Epidemiol. 62 (511–517), e511.
McAdam, K.G., Gregg, E.O., Bevan, M., Dittrich, D.J., Hemsley, S., Liu, C., Proctor, C.J., 2012. Design and chemical evaluation of reduced machine-yield cigarettes. Regul. Toxicol. Pharmacol. 62, 138–150.
McGreevy, K.M., Lipsitz, S.R., Linder, J.A., Rimm, E., Hoel, D.G., 2009. Using median regression to obtain adjusted estimates of central tendency for skewed laboratory and epidemiologic data. Clin. Chem. 55, 165–169.
Minet, E., Errington, G., Scherer, G., Newland, K., Sharifi, M., Bailey, B., McEwan, M., Cheung, F., 2011. An inter-laboratory comparison of urinary 3-hydroxypropylmercapturic acid measurement demonstrates good reproducibility between laboratories. BMC Res. Notes 4, 391.
Muggeo, V., 2013. quantregGrowth: Growth charts via regression quantiles. R package version 0.1-1. <http://cran.r-project.org/web/packages/quantregGrowth/> (Accessed: December 2014).
Muggeo, V., Sciandra, M., Tomasello, A., Calvo, S., 2013. Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. Environ. Ecol. Stat. 20, 519–531.
Noufaily, A., Jones, M., 2013. Parametric quantile regression based on the generalized gamma distribution. J. Royal Stat. Soc. C-App. 62, 723–740.
Peng, L., Wuu, J., Benatar, M., 2009. Developing reference data for nerve conduction studies: an application of quantile regression. Muscle Nerve 40, 763–771.
Perfetti, T., Rodgman, A., 2011. The complexity of tobacco and tobacco smoke. Beitr. Tabakforsch Int. 24, 215–232.
Piade, J.J., Wajrock, S., Jaccard, G., Janeke, G., 2013. Formation of mainstream cigarette smoke constituents prioritized by the World Health Organization–yield patterns observed in market surveys, clustering and inverse correlations. Food Chem. Toxicol. 55, 329–347.
Purkis, S., Intorp, M., 2014. Analysis of reference cigarette smoke yield data from 21 laboratories for 28 selected analytes as a guide to selection of new CORESTA recommended methods. Beitr. Tabakforsch Int. 26, 57–73.
Rodgman, A., Perfetti, T., 2009. The Chemical Components of Tobacco and Tobacco Smoke. CRC Press, Boca Raton, FL, USA.
SAS Institute Inc, Perfetti, T., 2011. SAS/STAT® 9.3 User's Guide. SAS Institute Inc., Cary, NC, USA.
Schall, R., 2012. The empirical coverage of confidence intervals: point estimates and confidence intervals for confidence levels. Biomed. J. 54, 537–551.
Soeteman-Hernandez, L.G., Bos, P.M., Talhout, R., 2013. Tobacco smoke-related health effects induced by 1,3-butadiene and strategies for risk reduction. Toxicol. Sci. 136, 566–580.
Tobacco Control Program Health Canada, 2004 Constituents and emissions reported for cigarettes sold in Canada – 2004.
US Department of Health and Human Services, 2001. Risks associated with smoking cigarettes with low machine-measured yields of tar and nicotine. National Institutes of Health, Bethesda, MD, USA.
US Food and Drug Administration, 2012. Harmful and potentially harmful constituents in tobacco products and tobacco smoke: established list. <http://www.fda.gov/TobaccoProducts/GuidanceComplianceRegulatoryInformation/ucm297786.htm> (Accessed: 18 March 2015).
Wei, Y., Pere, A., Koenker, R., He, X., 2006. Quantile regression methods for reference growth charts. Stat. Med. 25, 1369–1382.
WHO, 2005. WHO framework convention on tobacco control. World Health Organization, Geneva, Switzerland.